THE OHIO STATE UNIVERSITY

COLLEGE OF ENGINEERING

Stochastic Optimization for Large-scale Optimal Transport (OT)

Sixiong You

Mechanical and Aerospace Engineering Department

I. Introduction

- **Motivation**
- Optimal transport (OT) defines a powerful framework to compare probability distributions in a geometrically faithful way.
- Previous works are purely discrete and cannot cope with continuous densities, The only known class of methods that can overcome this limitation are so-called semi-discrete solvers.
- In addition, the practical impact of OT is still limited because of its computational burden
- This paper propose a new class of stochastic optimization algorithms to cope with large-scale OT problems.

I. Introduction

- This paper introduces three kinds of stochastic optimization methods to cope with three possible settings：
- **Discrete OT**: compare a discrete vs. another discrete measure

➢ Stochastic averaged gradient (SAG) method

- **Semi-discrete OT**: compare a discrete vs. a continuous measure

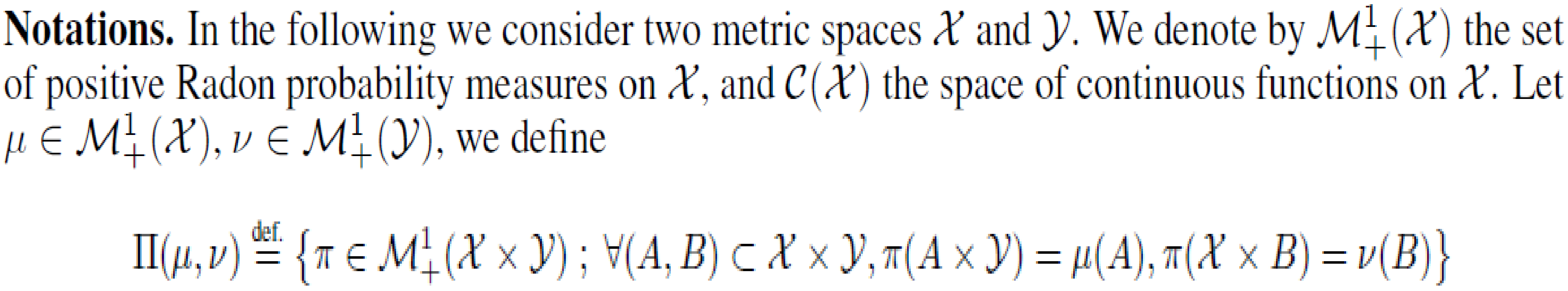
➢ Averaged stochastic gradient descent (SGD)

- **Continous OT**: to compare a continuous vs. another continuous measure

➢ makes use of an expansion of the dual variables in a reproducing kernel Hilbert space (RKHS)

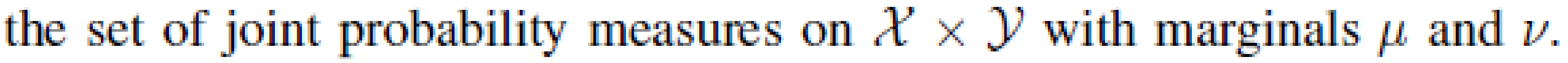
II. Problem Formulation

The definition of joint probability measures



Notations. In the following we consider two metric spaces $\mathcal{X}$ and $\mathcal{Y}$. We denote by $\mathcal{M}_+^1(\mathcal{X})$ the set of positive Radon probability measures on $\mathcal{X}$, and $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$. Let $\mu \in \mathcal{M}_+^1(\mathcal{X}), \nu \in \mathcal{M}_+^1(\mathcal{Y})$, we define

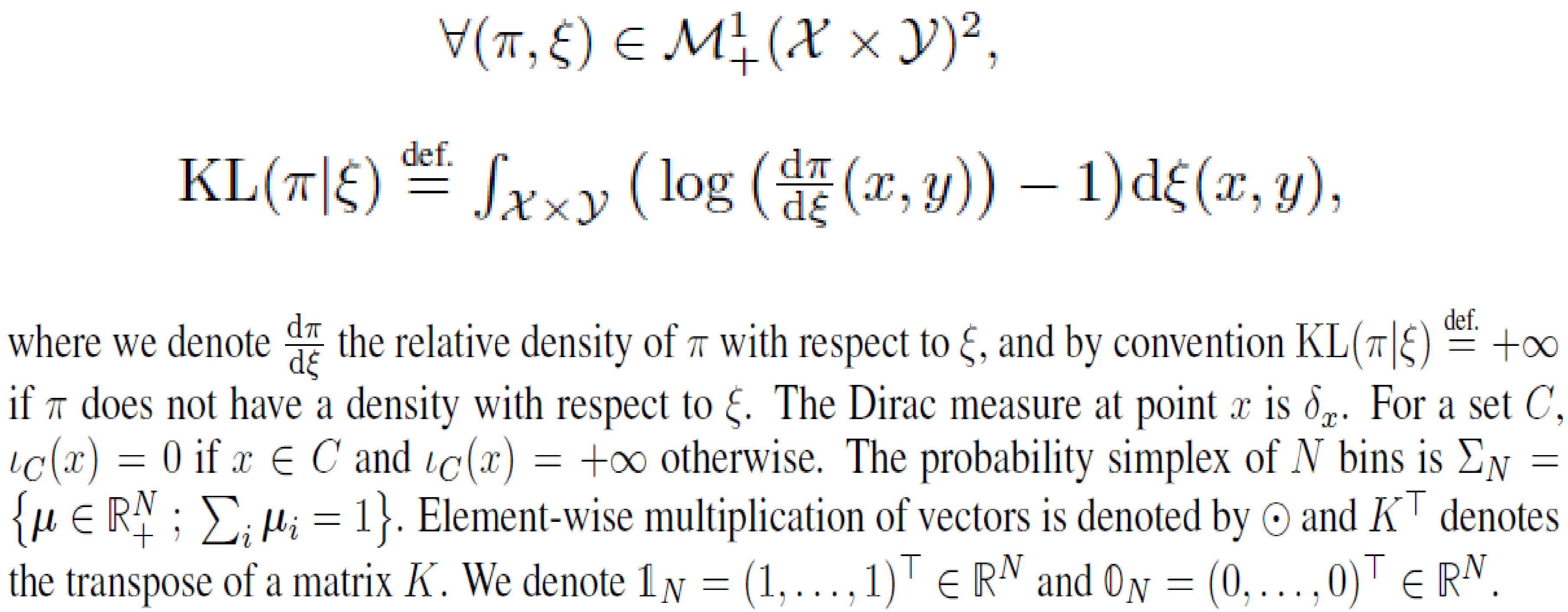
$$\Pi(\mu,\nu) \overset{\text{def.}}{=} \{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})\,;\, \forall (A,B) \subset \mathcal{X}\times\mathcal{Y}, \pi(A\times\mathcal{Y}) = \mu(A), \pi(\mathcal{X}\times B) = \nu(B)\}$$



the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

II. Problem Formulation

The definition of Kullback-Leibler divergence



$$\forall(\pi,\xi)\in\mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})^2,$$

$$\mathrm{KL}(\pi|\xi)\overset{\text{def.}}{=}\int_{\mathcal{X}\times\mathcal{Y}}\big(\log\big(\tfrac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\big)-1\big)\mathrm{d}\xi(x,y),$$

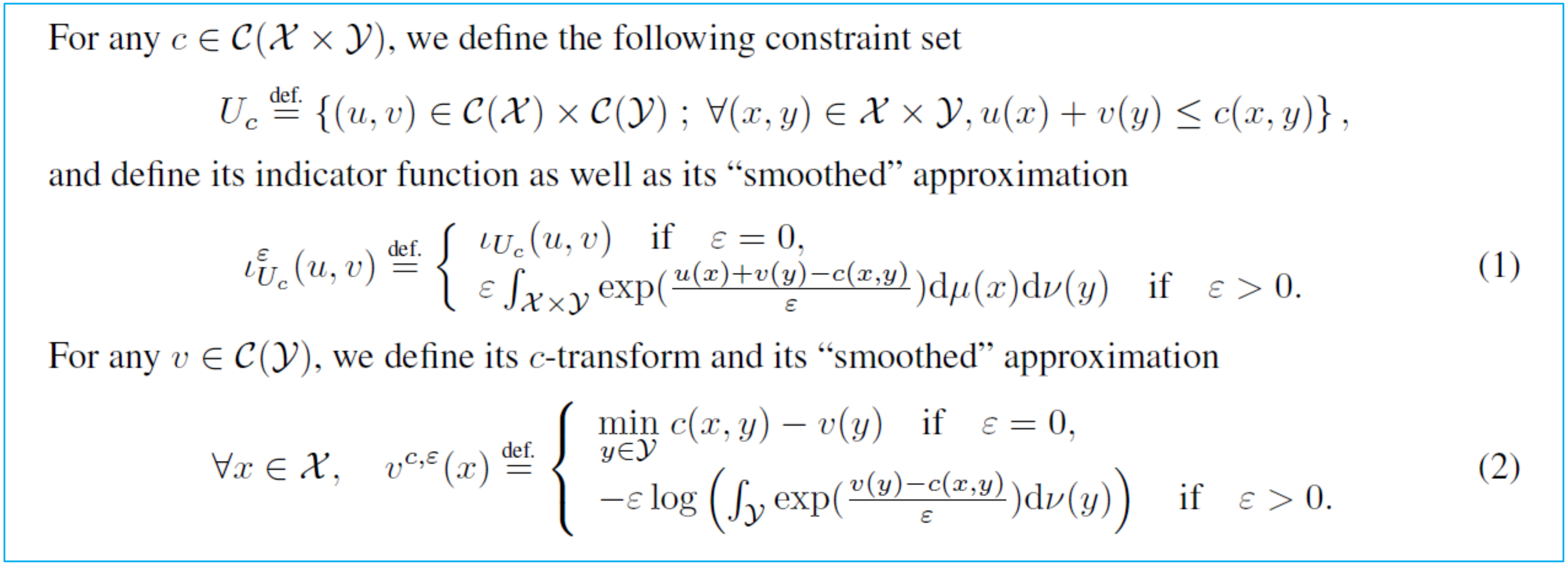
where we denote $\frac{\mathrm{d}\pi}{\mathrm{d}\xi}$ the relative density of π with respect to ξ, and by convention $\mathrm{KL}(\pi|\xi)\overset{\text{def.}}{=}+\infty$ if π does not have a density with respect to ξ. The Dirac measure at point x is δ_x. For a set C, $\iota_C(x)=0$ if $x\in C$ and $\iota_C(x)=+\infty$ otherwise. The probability simplex of N bins is $\Sigma_N=\{\mu\in\mathbb{R}_+^N\,;\,\sum_i\mu_i=1\}$. Element-wise multiplication of vectors is denoted by $\odot$ and $K^\top$ denotes the transpose of a matrix K. We denote $\mathbb{1}_N=(1,\ldots,1)^\top\in\mathbb{R}^N$ and $\mathbb{0}_N=(0,\ldots,0)^\top\in\mathbb{R}^N$.

II. Problem Formulation

The Kantorovich formulation of OT and its entropic regularization can be written in a single convex optimization problem

$$\forall(\mu,\nu)\in\mathcal{M}_+^1(\mathcal{X})\times\mathcal{M}_+^1(\mathcal{Y}),\ W_\varepsilon(\mu,\nu)\overset{\text{def.}}{=}\min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}}c(x,y)\mathrm{d}\pi(x,y)+\varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu).\quad(\mathcal{P}_\varepsilon)$$

In which $c(x,y)$ is the cost to move a unit of mass from x to y



For any $c\in\mathcal{C}(\mathcal{X}\times\mathcal{Y})$, we define the following constraint set

$$U_c\overset{\text{def.}}{=}\{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})\,;\,\forall(x,y)\in\mathcal{X}\times\mathcal{Y},u(x)+v(y)\leq c(x,y)\}\,,$$

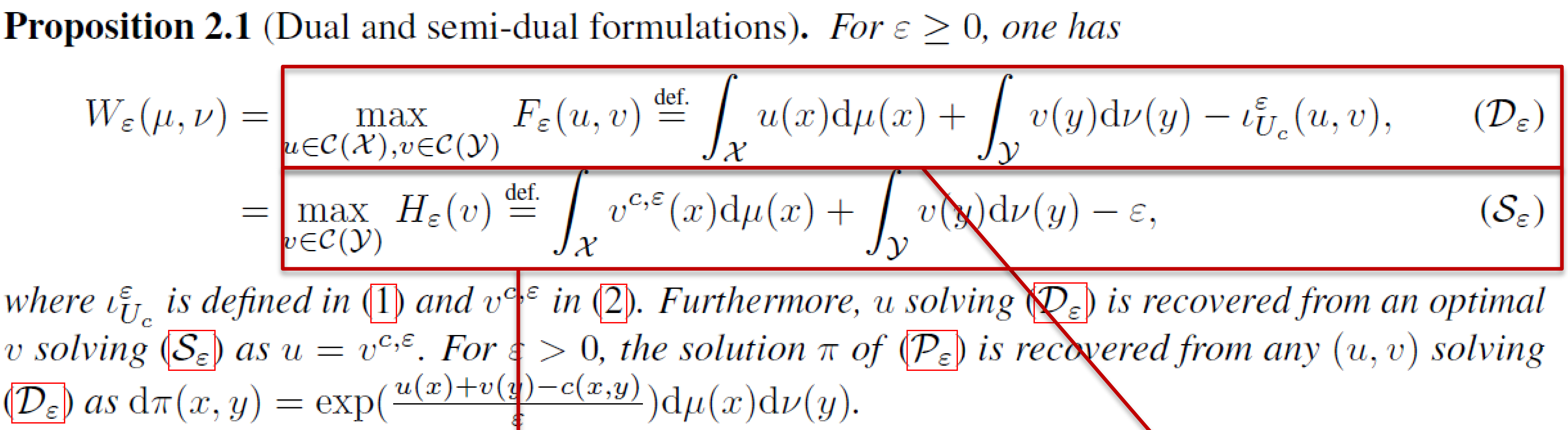
and define its indicator function as well as its “smoothed” approximation

$$\iota_{U_c}^{\varepsilon}(u,v)\overset{\text{def.}}{=}\begin{cases}\iota_{U_c}(u,v) & \text{if}\quad\varepsilon=0,\\ \varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\mu(x)\mathrm{d}\nu(y) & \text{if}\quad\varepsilon>0.\end{cases}\qquad(1)$$

For any $v\in\mathcal{C}(\mathcal{Y})$, we define its c-transform and its “smoothed” approximation

$$\forall x\in\mathcal{X},\quad v^{c,\varepsilon}(x)\overset{\text{def.}}{=}\begin{cases}\min_{y\in\mathcal{Y}}c(x,y)-v(y) & \text{if}\quad\varepsilon=0,\\ -\varepsilon\log\left(\int_{\mathcal{Y}}\exp(\frac{v(y)-c(x,y)}{\varepsilon})\mathrm{d}\nu(y)\right) & \text{if}\quad\varepsilon>0.\end{cases}\qquad(2)$$

II. Problem Formulation



Proposition 2.1 (Dual and semi-dual formulations). *For $\varepsilon \geq 0$, one has*

$$W_\varepsilon(\mu,\nu) = \max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} F_\varepsilon(u,v) \overset{\text{def.}}{=} \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(u,v), \qquad (\mathcal{D}_\varepsilon)$$

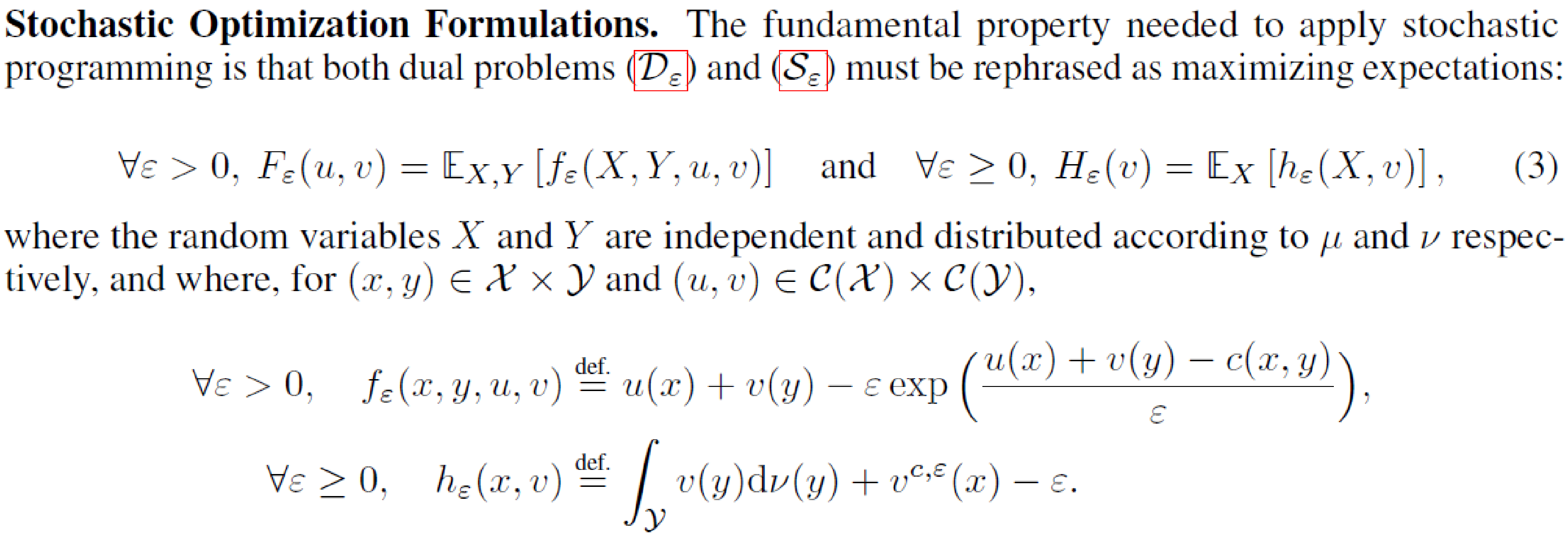
$$= \max_{v\in\mathcal{C}(\mathcal{Y})} H_\varepsilon(v) \overset{\text{def.}}{=} \int_{\mathcal{X}} v^{c,\varepsilon}(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \varepsilon, \qquad (\mathcal{S}_\varepsilon)$$

where $\iota^{\varepsilon}_{U_c}$ is defined in (1) and $v^{c,\varepsilon}$ in (2). Furthermore, u solving $(\mathcal{D}_\varepsilon)$ is recovered from an optimal v solving $(\mathcal{S}_\varepsilon)$ as $u = v^{c,\varepsilon}$. For $\varepsilon > 0$, the solution π of $(\mathcal{P}_\varepsilon)$ is recovered from any (u,v) solving $(\mathcal{D}_\varepsilon)$ as $\mathrm{d}\pi(x,y) = \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\mu(x)\mathrm{d}\nu(y)$.

Fenchel-Rockafellar's dual theorem

Solving $\varepsilon > 0, \frac{\partial F_\varepsilon(u,v)}{\partial u} = 0 \Rightarrow u = v^{c,\varepsilon}$,plugging this expression back in $D(\varepsilon)$

II. Problem Formulation



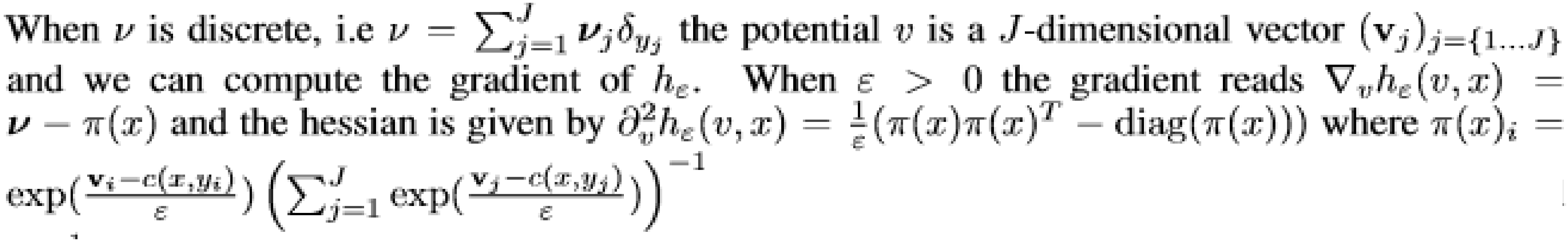
Stochastic Optimization Formulations. The fundamental property needed to apply stochastic programming is that both dual problems ($\boxed{\mathcal{D}_\varepsilon}$) and ($\boxed{\mathcal{S}_\varepsilon}$) must be rephrased as maximizing expectations:

$$\forall \varepsilon > 0,\ F_\varepsilon(u,v) = \mathbb{E}_{X,Y}\left[f_\varepsilon(X,Y,u,v)\right] \quad \text{and} \quad \forall \varepsilon \geq 0,\ H_\varepsilon(v) = \mathbb{E}_X\left[h_\varepsilon(X,v)\right], \qquad (3)$$

where the random variables X and Y are independent and distributed according to μ and ν respectively, and where, for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $(u,v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})$,

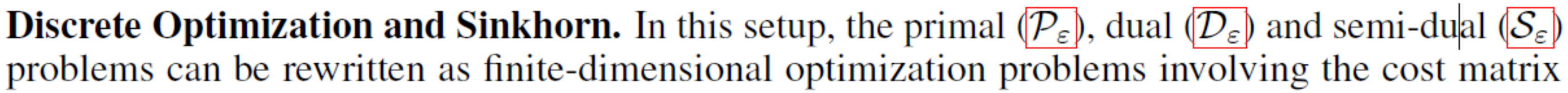
$$\forall \varepsilon > 0, \quad f_\varepsilon(x,y,u,v) \overset{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right),$$

$$\forall \varepsilon \geq 0, \quad h_\varepsilon(x,v) \overset{\text{def.}}{=} \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) + v^{c,\varepsilon}(x) - \varepsilon.$$

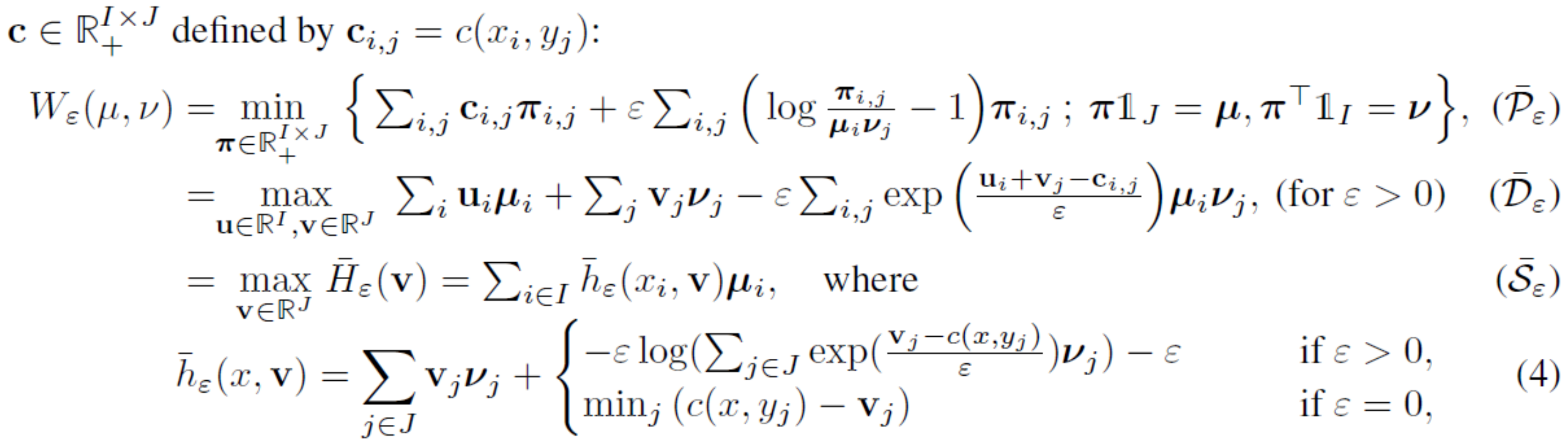


When ν is discrete, i.e $\nu = \sum_{j=1}^J \boldsymbol{\nu}_j \delta_{y_j}$ the potential v is a J-dimensional vector $(\mathbf{v}_j)_{j=\{1\ldots J\}}$ and we can compute the gradient of h_ε. When $\varepsilon > 0$ the gradient reads $\nabla_v h_\varepsilon(v,x) = \boldsymbol{\nu} - \pi(x)$ and the hessian is given by $\partial_v^2 h_\varepsilon(v,x) = \frac{1}{\varepsilon}(\pi(x)\pi(x)^T - \mathrm{diag}(\pi(x)))$ where $\pi(x)_i = \exp(\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon})\left(\sum_{j=1}^J \exp(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon})\right)^{-1}$

III. Discrete Optimal Transport



Discrete Optimization and Sinkhorn. In this setup, the primal ($\mathcal{P}_\varepsilon$), dual ($\mathcal{D}_\varepsilon$) and semi-dual ($\mathcal{S}_\varepsilon$) problems can be rewritten as finite-dimensional optimization problems involving the cost matrix



$\mathbf{c} \in \mathbb{R}_+^{I \times J}$ defined by $\mathbf{c}_{i,j} = c(x_i, y_j)$:

$$W_\varepsilon(\mu,\nu) = \min_{\boldsymbol{\pi} \in \mathbb{R}_+^{I \times J}} \left\{ \sum_{i,j} \mathbf{c}_{i,j} \boldsymbol{\pi}_{i,j} + \varepsilon \sum_{i,j} \left(\log \frac{\boldsymbol{\pi}_{i,j}}{\boldsymbol{\mu}_i \boldsymbol{\nu}_j} - 1 \right) \boldsymbol{\pi}_{i,j} \; ; \; \boldsymbol{\pi} \mathbb{1}_J = \boldsymbol{\mu}, \boldsymbol{\pi}^\top \mathbb{1}_I = \boldsymbol{\nu} \right\}, \quad (\bar{\mathcal{P}}_\varepsilon)$$

$$= \max_{\mathbf{u} \in \mathbb{R}^I, \mathbf{v} \in \mathbb{R}^J} \sum_i \mathbf{u}_i \boldsymbol{\mu}_i + \sum_j \mathbf{v}_j \boldsymbol{\nu}_j - \varepsilon \sum_{i,j} \exp\left(\frac{\mathbf{u}_i + \mathbf{v}_j - \mathbf{c}_{i,j}}{\varepsilon} \right) \boldsymbol{\mu}_i \boldsymbol{\nu}_j, \text{ (for } \varepsilon > 0) \quad (\bar{\mathcal{D}}_\varepsilon)$$

$$= \max_{\mathbf{v} \in \mathbb{R}^J} \bar{H}_\varepsilon(\mathbf{v}) = \sum_{i \in I} \bar{h}_\varepsilon(x_i, \mathbf{v}) \boldsymbol{\mu}_i, \quad \text{where} \quad (\bar{\mathcal{S}}_\varepsilon)$$

$$\bar{h}_\varepsilon(x, \mathbf{v}) = \sum_{j \in J} \mathbf{v}_j \boldsymbol{\nu}_j + \begin{cases} -\varepsilon \log\left(\sum_{j \in J} \exp\left(\frac{\mathbf{v}_j - c(x, y_j)}{\varepsilon}\right) \boldsymbol{\nu}_j\right) - \varepsilon & \text{if } \varepsilon > 0, \\ \min_j \left(c(x, y_j) - \mathbf{v}_j \right) & \text{if } \varepsilon = 0, \end{cases} \quad (4)$$

Stochastic gradient descent (SGD): the gradient of that term can be used as a proxy for the full gradient in a standard gradient ascent step to maximize

III. Discrete Optimal Transport

Stochastic gradient descent (SGD)

Example: For an optimization problem

Objective function $J = \min_{\omega} Q(\omega) = \min_{\omega} \frac{1}{n}\sum_{i=1}^{n} Q_i(\omega)$

When used to minimize the above function, a standard (or "batch") gradient descent method would perform the following iterations :

$$\omega := \omega - \eta \nabla Q(\omega) = \omega - \eta \sum_{i=1}^{n} \nabla Q(\omega)/n$$

where η is a step size. However, evaluating the sum-gradient may require expensive evaluations of the gradients from all summand functions. To economize on the computational cost at every iteration, stochastic gradient descent samples a subset of summand functions at every step. This is very effective in the case of large-scale machine learning problems.

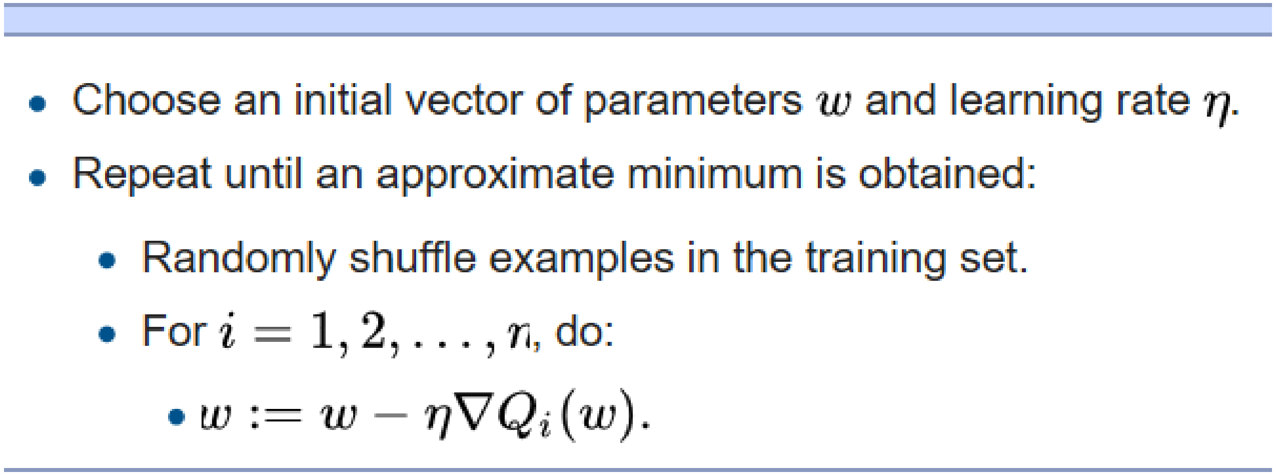
III. Discrete Optimal Transport

Stochastic gradient descent (SGD)

In stochastic (or "on-line") gradient descent, the true gradient of $Q(\omega)$ is approximated by a gradient at a single example

$$\omega := \omega - \eta \nabla Q_i(\omega)$$

In pseudocode, stochastic gradient descent can be presented as follows:



```
• Choose an initial vector of parameters $w$ and learning rate $\eta$.
• Repeat until an approximate minimum is obtained:
    • Randomly shuffle examples in the training set.
    • For $i = 1, 2, \ldots, n$, do:
        • $w := w - \eta \nabla Q_i(w)$.
```

III. Discrete Optimal Transport

Averaged stochastic gradient descent (Average SGD)

Invented independently by Ruppert and Polyak in the late 1980s, is ordinary stochastic gradient descent that records an average of its parameter vector over time. That is, the update is the same as for ordinary stochastic gradient descent, but the algorithm also keeps track of

$$\bar{\omega} = \frac{1}{t}\sum_{i=0}^{t-1}\omega_i$$

When optimization is done, this averaged parameter vector takes the place of ω

Stochastic averaged gradient (SAG): the stochastic average gradient method with a (user-supplied) constant step size.

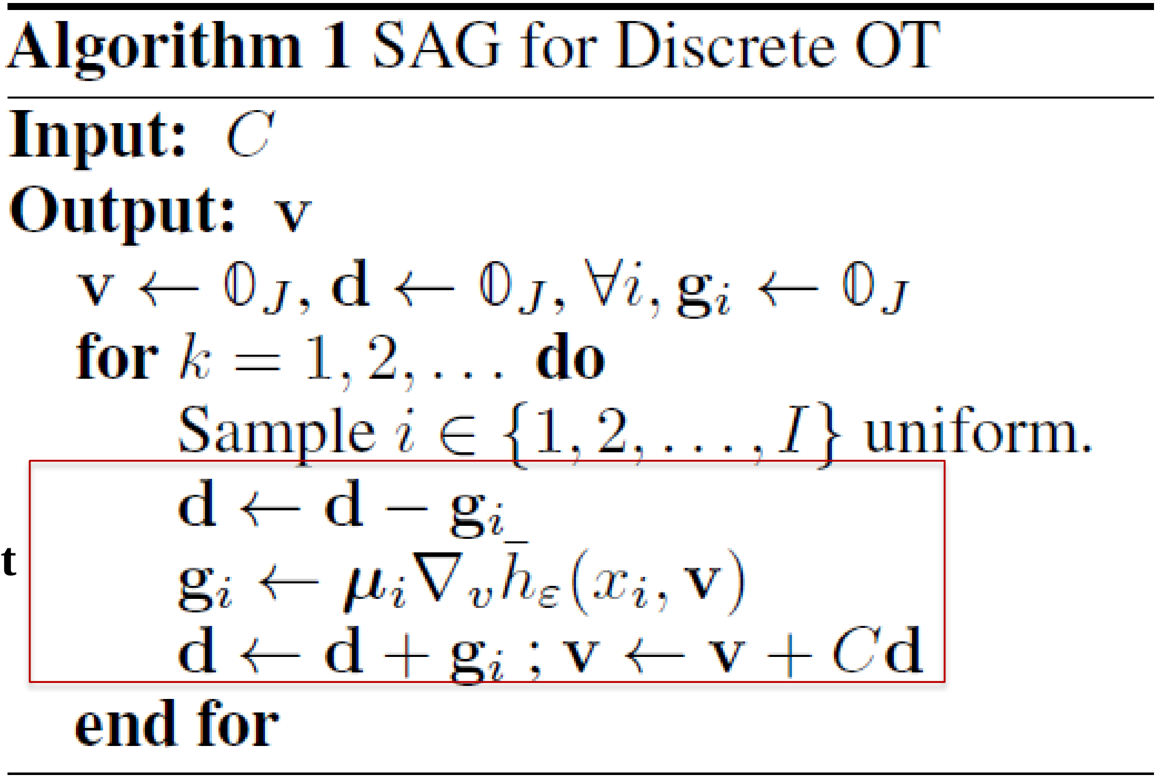
III. Discrete Optimal Transport

The flow chat of SAG for discrete OT

Stepsize

Output

Initial g



```
Algorithm 1 SAG for Discrete OT
Input: $C$
Output: $\mathbf{v}$
    $\mathbf{v} \leftarrow \mathbb{0}_J, \mathbf{d} \leftarrow \mathbb{0}_J, \forall i, \mathbf{g}_i \leftarrow \mathbb{0}_J$
    for $k = 1, 2, \ldots$ do
        Sample $i \in \{1, 2, \ldots, I\}$ uniform.
        $\mathbf{d} \leftarrow \mathbf{d} - \mathbf{g}_i$
        $\mathbf{g}_i \leftarrow \mu_i \nabla_v \bar{h}_\varepsilon(x_i, \mathbf{v})$
        $\mathbf{d} \leftarrow \mathbf{d} + \mathbf{g}_i$ ; $\mathbf{v} \leftarrow \mathbf{v} + C\mathbf{d}$
    end for
```

Update gradient

III. Discrete Optimal Transport

Numerical Illustrations on Bags of Word-Embeddings

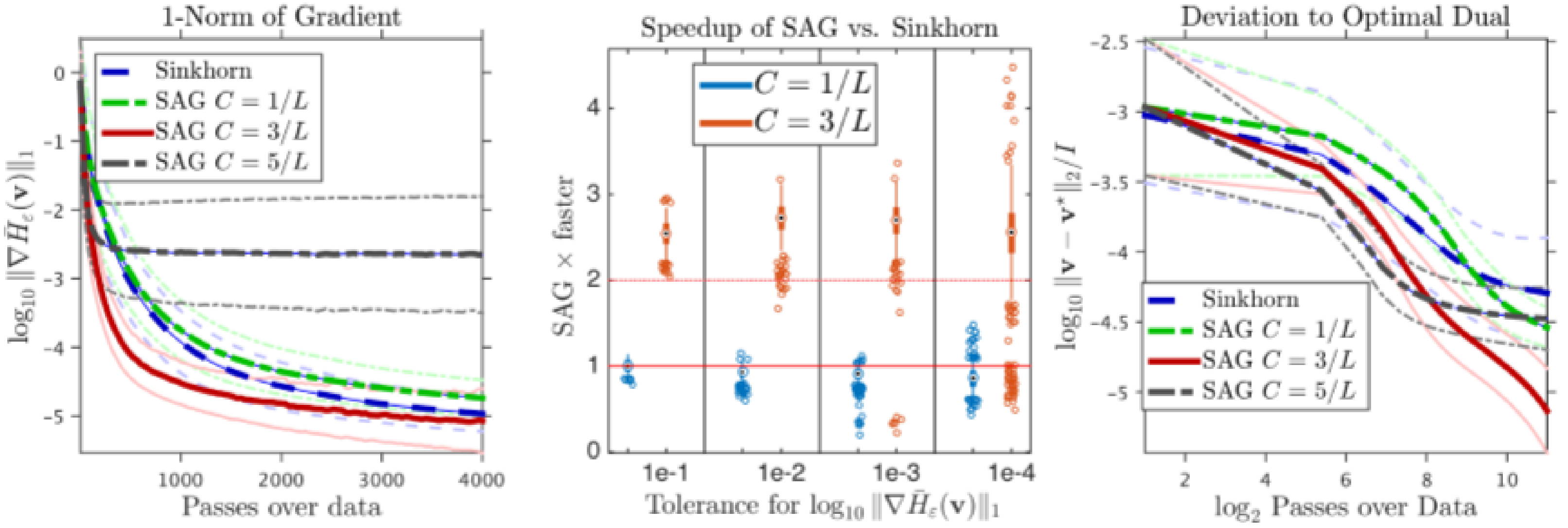
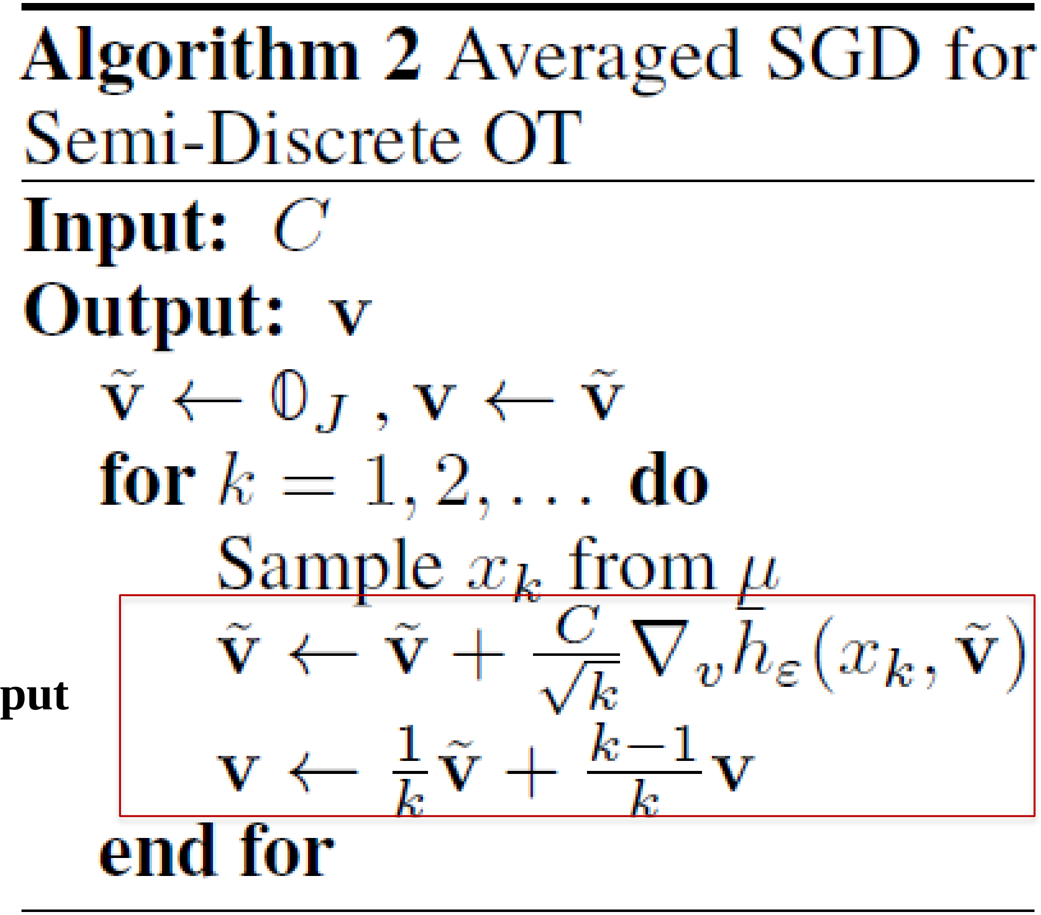


Figure 1: We compute all 595 pairwise word mover's distances [11] between 35 very large corpora of text, each represented as a cloud of $I = 20,000$ word embeddings. We compare the Sinkhorn algorithm with SAG, tuned with different stepsizes. Each pass corresponds to a $I \times I$ matrix-vector product. We used minibatches of size 200 for SAG. *Left plot*: convergence of the gradient ℓ_1 norm (average and $\pm$ standard deviation error bars). A stepsize of $3/L$ achieves a substantial speed-up of ≈ 2.5, as illustrated in the boxplots in the *center plot*. Convergence to $\mathbf{v}^\star$ (the best dual variable across all variables after $4,000$ passes) in ℓ_2 norm is given in the *right plot*, up to $2,000 \approx 2^{11}$ steps.

IV. Semi-Discrete Optimal Transport

The flow chat of SGD for discrete OT



Algorithm 2 Averaged SGD for Semi-Discrete OT

```
Input: $C$
Output: $\mathbf{v}$
    $\tilde{\mathbf{v}} \leftarrow \mathbb{0}_J$ , $\mathbf{v} \leftarrow \tilde{\mathbf{v}}$
    for $k = 1, 2, \ldots$ do
        Sample $x_k$ from $\mu$
        $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} + \frac{C}{\sqrt{k}} \nabla_v h_\varepsilon(x_k, \tilde{\mathbf{v}})$
        $\mathbf{v} \leftarrow \frac{1}{k} \tilde{\mathbf{v}} + \frac{k-1}{k} \mathbf{v}$
    end for
```

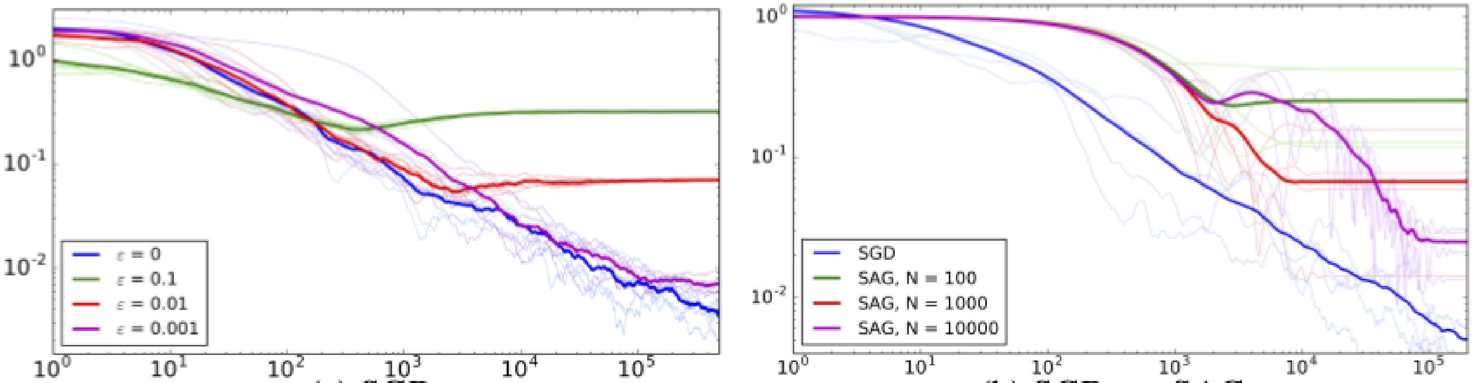
Stepsize

Output

Update output

IV. Semi-Discrete Optimal Transport

Numerical Illustrations

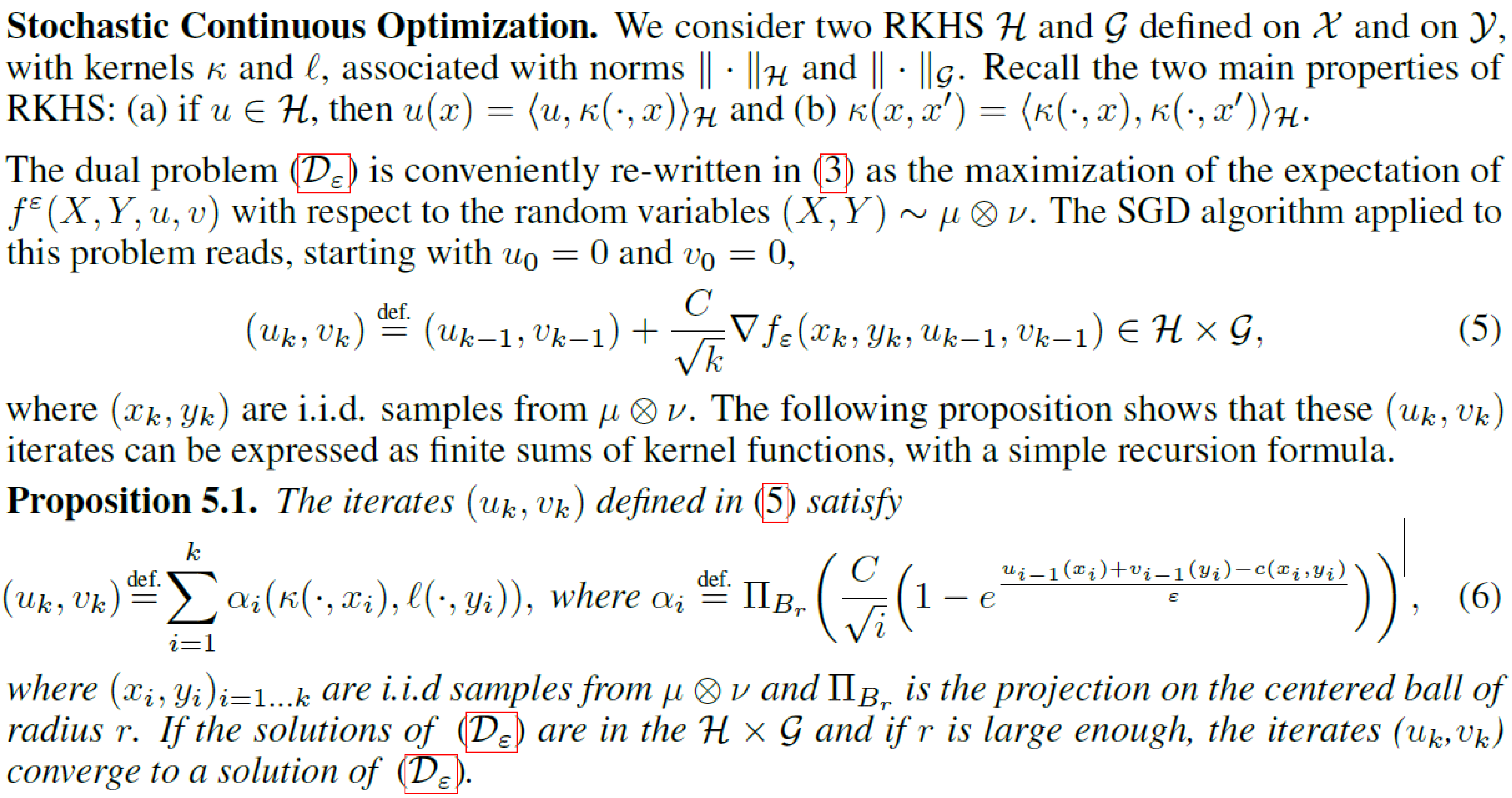


(a) SGD (b) SGD vs. SAG

Figure 2: (a) Plot of $\|\mathbf{v}_k - \mathbf{v}_0^\star\|_2 \,/\, \|\mathbf{v}_0^\star\|_2$ as a function of k, for SGD and different values of ε ($\varepsilon = 0$ being un-regularized). (b) Plot of $\|\mathbf{v}_k - \mathbf{v}_\varepsilon^\star\|_2 \,/\, \|\mathbf{v}_\varepsilon^\star\|_2$ as a function of k, for SGD and SAG with different number N of samples, for regularized OT using $\varepsilon = 10^{-2}$.

Figure 2 (a) shows the evolution of $\|\mathbf{v}_k - \mathbf{v}_0^\star\|_2 \,/\, \|\mathbf{v}_0^\star\|_2$ as a function of k. It highlights the influence of the regularization parameters ε on the iterates of SGD. While the regularized iterates converge faster, they do not converge to the correct unregularized solution. This figure also illustrates the convergence theorem of solution of $(\mathcal{S}_\varepsilon)$ toward those $(\mathcal{S}_0)$ when $\varepsilon \to 0$, which can be found in the supplementary material. Figure 2 (b) shows the evolution of $\|\mathbf{v}_k - \mathbf{v}_\varepsilon^\star\|_2 \,/\, \|\mathbf{v}_\varepsilon^\star\|_2$ as a function of k, for a fixed regularization parameter value $\varepsilon = 10^{-2}$. It compares SGD to SAG using different numbers N of samples for the empirical measures $\hat{\mu}_N$. While SGD converges to the true solution of the semi-discrete problem, the solution computed by SAG is biased because of the approximation error which comes from the discretization of μ. This error decreases when the sample size N is increased, as the approximation of μ by $\hat{\mu}_N$ becomes more accurate.

IV. Continuous Optimal Transport



Stochastic Continuous Optimization. We consider two RKHS $\mathcal{H}$ and $\mathcal{G}$ defined on $\mathcal{X}$ and on $\mathcal{Y}$, with kernels κ and ℓ, associated with norms $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{G}}$. Recall the two main properties of RKHS: (a) if $u \in \mathcal{H}$, then $u(x) = \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}}$ and (b) $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x')\rangle_{\mathcal{H}}$.

The dual problem ($\mathcal{D}_\varepsilon$) is conveniently re-written in (3) as the maximization of the expectation of $f^\varepsilon(X, Y, u, v)$ with respect to the random variables $(X, Y) \sim \mu \otimes \nu$. The SGD algorithm applied to this problem reads, starting with $u_0 = 0$ and $v_0 = 0$,

$$(u_k, v_k) \overset{\text{def.}}{=} (u_{k-1}, v_{k-1}) + \frac{C}{\sqrt{k}} \nabla f_\varepsilon(x_k, y_k, u_{k-1}, v_{k-1}) \in \mathcal{H} \times \mathcal{G}, \tag{5}$$

where (x_k, y_k) are i.i.d. samples from $\mu \otimes \nu$. The following proposition shows that these (u_k, v_k) iterates can be expressed as finite sums of kernel functions, with a simple recursion formula.

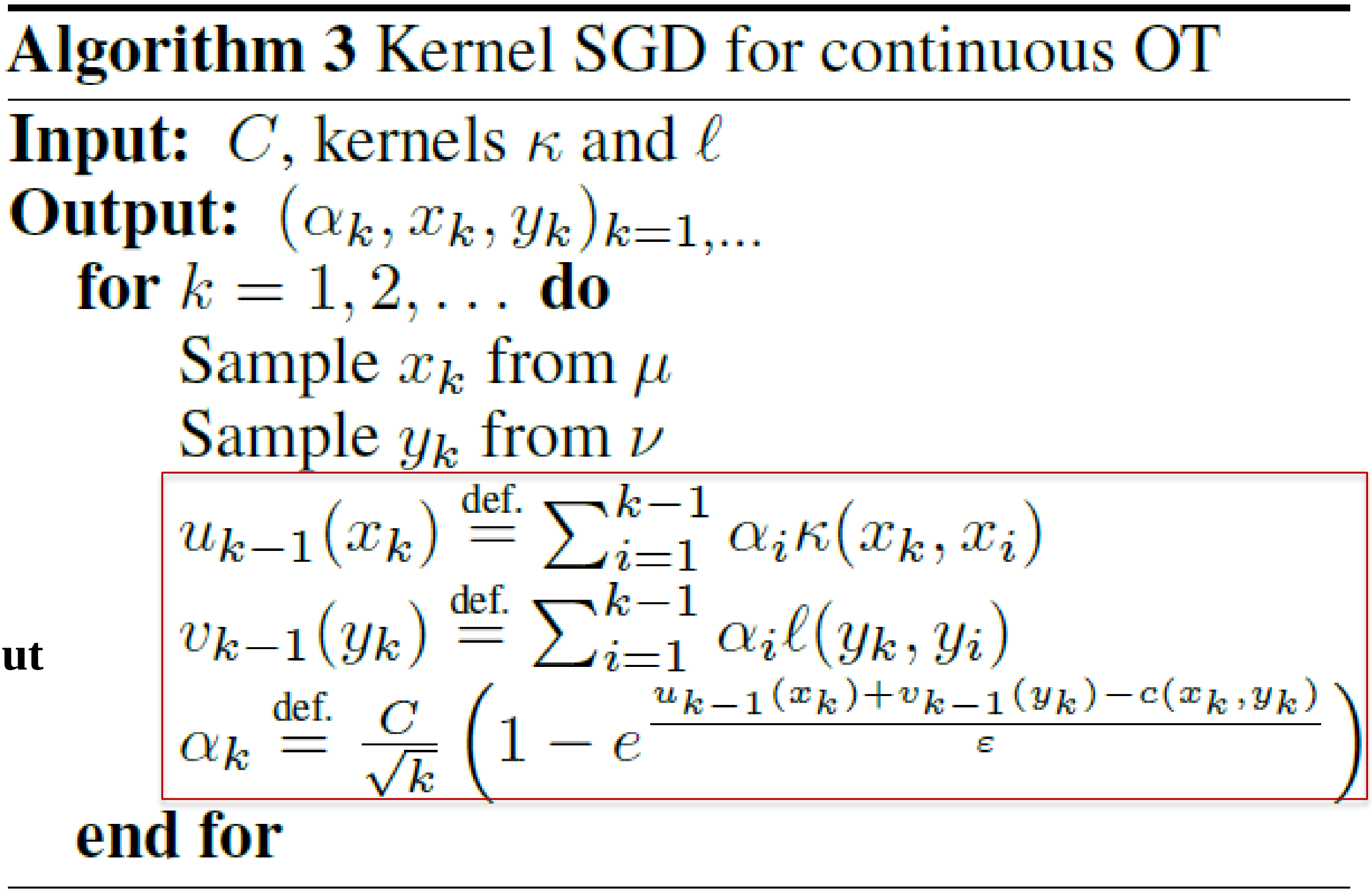
Proposition 5.1. *The iterates (u_k, v_k) defined in (5) satisfy*

$$(u_k, v_k) \overset{\text{def.}}{=} \sum_{i=1}^{k} \alpha_i(\kappa(\cdot, x_i), \ell(\cdot, y_i)), \text{ where } \alpha_i \overset{\text{def.}}{=} \Pi_{B_r}\left(\frac{C}{\sqrt{i}}\left(1 - e^{\frac{u_{i-1}(x_i)+v_{i-1}(y_i)-c(x_i,y_i)}{\varepsilon}}\right)\right), \tag{6}$$

where $(x_i, y_i)_{i=1\ldots k}$ are i.i.d samples from $\mu \otimes \nu$ and Π_{B_r} is the projection on the centered ball of radius r. If the solutions of ($\mathcal{D}_\varepsilon$) are in the $\mathcal{H} \times \mathcal{G}$ and if r is large enough, the iterates (u_k, v_k) converge to a solution of ($\mathcal{D}_\varepsilon$).

IV. Continuous Optimal Transport

The flow chat of Kernel SGD for discrete OT



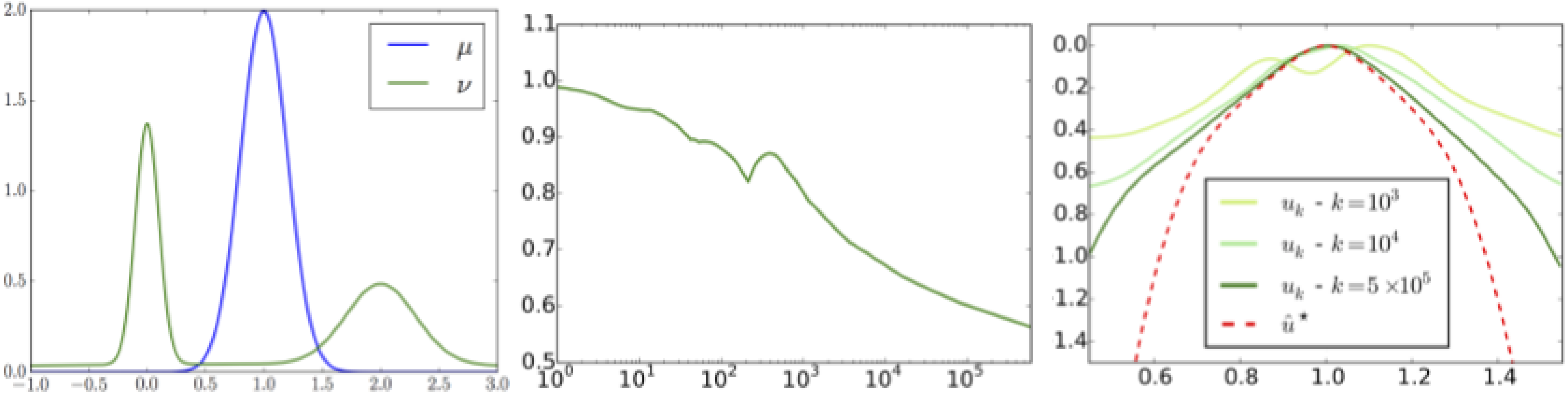
```
Algorithm 3 Kernel SGD for continuous OT
Input: $C$, kernels $\kappa$ and $\ell$
Output: $(\alpha_k, x_k, y_k)_{k=1,\dots}$
  for $k = 1, 2, \dots$ do
      Sample $x_k$ from $\mu$
      Sample $y_k$ from $\nu$
      $u_{k-1}(x_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} \alpha_i \kappa(x_k, x_i)$
      $v_{k-1}(y_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} \alpha_i \ell(y_k, y_i)$
      $\alpha_k \overset{\text{def.}}{=} \frac{C}{\sqrt{k}} \left(1 - e^{\frac{u_{k-1}(x_k) + v_{k-1}(y_k) - c(x_k, y_k)}{\varepsilon}}\right)$
  end for
```

Stepsize C and Kernels

Update output

IV. Continuous Optimal Transport

Numerical Illustrations



(a) setting

(b) convergence of u_k

(c) plots of u_k

Figure 3: (a) Plot of $\frac{\mathrm{d}\mu}{\mathrm{d}x}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}x}$. (b) Plot of $\|\mathbf{u}_k - \hat{\mathbf{u}}^\star\|_2 \,/\, \|\hat{\mathbf{u}}^\star\|_2$ as a function of k with SGD in the RKHS, for regularized OT using $\varepsilon = 10^{-1}$. (c) Plot of the iterates u_k for $k = 10^3, 10^4, 10^5$ and the proxy for the true potential $\hat{\mathbf{u}}^\star$, evaluated on a grid where μ has non negligible mass.

Thank You!