Beyond Euclidean Gossip: KL-Barycentric Consensus on Heterogeneous and Imbalanced Images

Supplementary Material

A. Deriving the closed-form prior to the local natural gradient.

We give the details of the local natural gradient in Eq. (5). Recall the local ELBO term can be written as

$$\mathcal{L}_i(\lambda) = w_i \, \mathbb{E}_{q_\lambda}[\log p(\theta) - \log q_\lambda(\theta)] + \mathbb{E}_{q_\lambda}[\log p(X_i|\theta)] \,,$$

where q_λ is a (minimal) exponential-family variational posterior with log-partition $A(\lambda)$, and $p(\theta)$ is the conjugate EF prior with natural parameter η_0. We focus on the first term, since it admits a closed form in λ.

Step 1 (rewrite the EF KL/prior term in canonical form). For a conjugate EF pair, the difference of log-densities satisfies

$$\log p(\theta) - \log q_\lambda(\theta) = \langle \phi(\theta), \eta_0 - \lambda \rangle - \big(A(\eta_0) - A(\lambda)\big),$$

hence taking expectation under q_λ gives

$$\begin{aligned} \mathbb{E}_{q_\lambda}\left[\log \frac{p(\theta)}{q_\lambda(\theta)}\right] &= \big\langle \mathbb{E}_{q_\lambda}[\phi(\theta)],\, \eta_0 - \lambda \big\rangle - \big(A(\eta_0) - A(\lambda)\big) \\ &= \langle \nabla A(\lambda),\, \eta_0 - \lambda \rangle + A(\lambda) - A(\eta_0), \end{aligned}$$

where we used the EF moment identity $\mathbb{E}_{q_\lambda}[\phi(\theta)] = \nabla A(\lambda)$. The constant $-A(\eta_0)$ does not affect gradients and will be dropped below.

Step 2 (compute the Euclidean gradient w.r.t. λ). Differentiate the non-constant part $\langle \nabla A(\lambda), \eta_0 - \lambda \rangle + A(\lambda)$ using the product rule,

$$\nabla_\lambda\Big(\langle \nabla A(\lambda), \eta_0 - \lambda \rangle\Big) = \nabla^2 A(\lambda)\,(\eta_0 - \lambda) \;-\; \nabla A(\lambda), \tag{9}$$

$$\nabla_\lambda A(\lambda) = \nabla A(\lambda). \tag{10}$$

Adding (9) and (10), the $\pm \nabla A(\lambda)$ terms cancel, yielding

$$\nabla_\lambda \, \mathbb{E}_{q_\lambda}\left[\log \frac{p(\theta)}{q_\lambda(\theta)}\right] = \nabla^2 A(\lambda)\,(\eta_0 - \lambda). \tag{11}$$

Step 3 (convert to the natural gradient). By definition, $\widetilde{\nabla}_\lambda \mathcal{L}_i(\lambda) = F(\lambda)^{-1} \nabla_\lambda \mathcal{L}_i(\lambda)$. For a minimal EF, the Fisher information of q_λ equals the Hessian of the log-partition,

$$F(\lambda) = \nabla^2 A(\lambda). \tag{12}$$

Combining (11) and (12) gives the closed-form identity,

$$\begin{aligned} F(\lambda)^{-1} \, \nabla_\lambda \, \mathbb{E}_{q_\lambda}\left[\log \frac{p(\theta)}{q_\lambda(\theta)}\right] &= F(\lambda)^{-1} \, F(\lambda)\,(\eta_0 - \lambda) \\ &= \eta_0 - \lambda. \end{aligned}$$

Thus, the prior part contributes exactly an additive term $(\eta_0 - \lambda)$ to the natural gradient (scaled by w_i if we split the KL regularizer across clients).

Step 4 (assemble the local natural gradient). The likelihood term $\mathbb{E}_{q_\lambda}[\log p(X_i|\theta)]$ typically has no closed form for deep models, so we keep its natural gradient as-is (estimated by minibatches and backprop through the variational moments). Altogether, we have

$$\widetilde{\nabla}_\lambda \mathcal{L}_i(\lambda_{i,t}) = w_i(\eta_0 - \lambda_{i,t}) + \widetilde{\nabla}_{\lambda_{i,t}} \, \mathbb{E}_{q_{\lambda_{i,t}}}[\log p(X_i|\theta)] \,.$$

Remarks. (i) The cancellation in (11) is the key reason conjugate EF VI admits simple natural-gradient updates: the KL/prior term becomes linear in $(\eta_0 - \lambda)$ after preconditioning by $F(\lambda)^{-1}$. (ii) Equality $F(\lambda) = \nabla^2 A(\lambda)$ in (12) holds when the EF representation is minimal; otherwise F may be singular and one typically works in a minimal parameterization or uses a pseudoinverse. (iii) This derivation uses only EF identities of the variational family q_λ; it does not assume the deep network likelihood is an EF model.

B. Proof of Theorem 1

We give a self-contained proof sketch with explicit disagreement control and the resulting optimization rates.

B.1. Preliminaries: EF duality and mean-space dynamics

We consider a minimal exponential-family variational posterior with log-partition A. Let λ be the natural parameter and $\mu = \nabla A(\lambda)$ be the expectation parameter. The Fisher information is $F(\lambda) = \nabla^2 A(\lambda)$. By EF duality, the natural gradient in λ coincides with an ordinary gradient in μ,

$$\tilde{\nabla}_\lambda \mathcal{L}(\lambda) = F(\lambda)^{-1} \nabla_\lambda \mathcal{L}(\lambda) = \nabla_\mu \mathcal{L}^*(\mu), \quad \mu = \nabla A(\lambda).$$

We assume metric well-conditioning on the iterate set: there exist constants $0 < \ell_F \le u_F < \infty$ such that

$$\ell_F I \preceq F(\lambda) \preceq u_F I, \quad \text{for all iterates } \lambda. \tag{13}$$

Equivalently, A is ℓ_F-strongly convex and u_F-smooth, and the mirror map $\mu = \nabla A(\lambda)$ is u_F-Lipschitz and ℓ_F-strongly monotone.

Quadratic bounds for the EF Bregman divergence. Let $D_A(\lambda, \lambda') = A(\lambda) - A(\lambda') - \langle \nabla A(\lambda'), \lambda - \lambda' \rangle$. By Eq. (13)

and Bregman duality $D_A(\lambda, \lambda') = D_{A^*}(\mu', \mu)$ with $\mu = \nabla A(\lambda)$, $\mu' = \nabla A(\lambda')$ [4, 5], we have for all iterates,

$$\frac{\ell_F}{2}\,\|\mu-\nu\|_2^2 \;\le\; D_A\big(\lambda(\mu),\lambda(\nu)\big) \;\le\; \frac{u_F}{2}\,\|\mu-\nu\|_2^2. \quad (14)$$

Mean-space decentralized recursion. Let M be the number of clients and we stack $\mu_t = [\mu_{1,t}; \dots; \mu_{M,t}] \in \mathbb{R}^{Md}$. Let $W = [w_{ij}]$ be symmetric, doubly stochastic, and let $J = \frac{1}{M}\mathbf{1}\mathbf{1}^\top$. The KL-consensus update yields a descent-and-mixing recursion in μ of the form,

$$\mu_{t+1} = (W \otimes I)\mu_t - \rho_t\, g_t, \quad (15)$$

where $g_t = [g_{1,t}; \dots; g_{M,t}]$ denotes the (stochastic) mean-space gradients corresponding to $\nabla f_i(\mu_{i,t})$ (and any linear "prior pull" term, which is Lipschitz under Eq. (13) and can be absorbed into $g_{i,t}$). Define $f(\mu) = \sum_{i=1}^M f_i(\mu)$ as in the main text, and let $\bar\mu_t = \frac{1}{M}\sum_{i=1}^M \mu_{i,t}$.

Regularity (standard). Assume f is L-smooth in μ. Assume $g_{i,t}$ is conditionally unbiased with bounded second moment:

$$\mathbb{E}[g_{i,t}|\mathcal{F}_t] = \nabla f_i(\mu_{i,t}), \quad \mathbb{E}\big[\|g_{i,t}\|_2^2|\mathcal{F}_t\big] \le G^2, \quad (16)$$

where $\mathcal{F}_t$ is the filtration up to time t [7, 17].

B.2. Consensus contraction and controlled client drift

Let the disagreement (consensus error) be

$$E_t = \sum_{i=1}^M \|\mu_{i,t} - \bar\mu_t\|_2^2 = \|(I-J)\mu_t\|_2^2.$$

Let $\Delta = \|W - J\|_2 < 1$ be the disagreement contraction factor (equivalently, spectral gap $1-\Delta$).

Lemma 1 (Disagreement recursion)**.** *Under Eqs.* (15)–(16)*,*

$$\mathbb{E}[E_{t+1}|\mathcal{F}_t] \le \Delta^2 E_t + C_1\,\rho_t^2, \quad C_1 \le c\,MG^2, \quad (17)$$

for an absolute numerical constant $c > 0$. Consequently,

$$\mathbb{E}[E_t] \le \Delta^{2t}E_0 + C_1\sum_{k=0}^{t-1}\Delta^{2(t-1-k)}\rho_k^2 = \mathcal{O}\Big(\tfrac{1}{1-\Delta}\max_{k<t}\rho_k^2\Big).$$

In the full-batch case ($G = 0$), $\mathbb{E}[E_{t+1}] \le \Delta^2\mathbb{E}[E_t]$.

Proof. Using $JW = WJ = J$ and Eq. (15),

$$(I-J)\mu_{t+1} = (W-J)\mu_t - \rho_t(I-J)g_t.$$

Taking norms and using $\|(W-J)x\|_2 \le \Delta\|x\|_2$, we obtain

$$E_{t+1} \le \Delta^2 E_t + 2\rho_t^2\|(I-J)g_t\|_2^2$$

after a standard quadratic expansion and absorbing the cross term by Young's inequality. Moreover, $\|I-J\|_2 \le 1$ and $\mathbb{E}[\|(I-J)g_t\|_2^2|\mathcal{F}_t] \le \sum_{i=1}^M \mathbb{E}[\|g_{i,t}\|_2^2|\mathcal{F}_t] \le MG^2$, which yields Eq. (17). Iterating gives the stated bound. □

B.3. Averaged dynamics and gradient perturbation

By double stochasticity, averaging Eq. (15) yields

$$\bar\mu_{t+1} = \bar\mu_t - \rho_t\,\bar g_t, \qquad \bar g_t = \tfrac{1}{M}\sum_{i=1}^M g_{i,t}. \quad (18)$$

We introduce the *virtual centralized* gradient step,

$$\tilde\mu_{t+1} = \bar\mu_t - \rho_t\,\nabla f(\bar\mu_t). \quad (19)$$

Lemma 2 (Gradient mismatch is controlled by disagreement)**.** *If f_i is L-smooth, then*

$$\Big\|\tfrac{1}{M}\sum_{i=1}^M \nabla f_i(\mu_{i,t}) - \nabla f(\bar\mu_t)\Big\|_2 \le L\sqrt{\tfrac{E_t}{M}}. \quad (20)$$

Proof. Using Jensen and Lipschitz gradients,

$$\begin{aligned}
&\Big\|\tfrac{1}{M}\sum_{i=1}^M(\nabla f_i(\mu_{i,t}) - \nabla f_i(\bar\mu_t))\Big\| \\
\le&\tfrac{1}{M}\sum_{i=1}^M L\|\mu_{i,t} - \bar\mu_t\| \\
\le&L\sqrt{\tfrac{1}{M}\sum_{i=1}^M\|\mu_{i,t} - \bar\mu_t\|^2} \\
=&L\sqrt{\tfrac{E_t}{M}}.
\end{aligned}$$

□

Lemma 3 (Average iterate tracks the virtual centralized step)**.** *Under Eqs.* (16) *and* (20)*,*

$$\mathbb{E}\big[\|\bar\mu_{t+1} - \tilde\mu_{t+1}\|_2^2\,\big|\,\mathcal{F}_t\big] \le 2\rho_t^2L^2\tfrac{E_t}{M} + 2\rho_t^2\tfrac{G^2}{M}.$$

Proof. From Eqs. (18)–(19), we have

$$\bar\mu_{t+1} - \tilde\mu_{t+1} = -\rho_t\Big(\bar g_t - \nabla f(\bar\mu_t)\Big).$$

We add and subtract $\frac{1}{M}\sum_{i=1}^M\nabla f_i(\mu_{i,t})$, then apply $(a+b)^2 \le 2a^2 + 2b^2$, use Lemma 2 for the bias term, and Eq. (16) for the variance term,

$$\mathbb{E}\left[\Big\|\tfrac{1}{M}\sum_{i=1}^M(g_{i,t} - \nabla f_i(\mu_{i,t}))\Big\|^2\Big|\mathcal{F}_t\right] \le G^2/M$$

. □

B.4. Descent recursion

By L-smoothness of f, the virtual step Eq. (19) satisfies

$$f(\tilde\mu_{t+1}) \le f(\bar\mu_t) - \rho_t\|\nabla f(\bar\mu_t)\|_2^2 + \frac{L}{2}\rho_t^2\|\nabla f(\bar\mu_t)\|_2^2.$$

Also by L-smoothness, $f(\bar{\mu}_{t+1}) \le f(\tilde{\mu}_{t+1}) + \frac{L}{2}\|\bar{\mu}_{t+1} - \tilde{\mu}_{t+1}\|_2^2$. Combining with Lemma 3 and taking conditional expectation yields

$$\mathbb{E}[f(\bar{\mu}_{t+1})|\mathcal{F}_t] \le f(\bar{\mu}_t) - \rho_t\Big(1 - \tfrac{L\rho_t}{2}\Big)\|\nabla f(\bar{\mu}_t)\|_2^2 + L\rho_t^2\Big(L^2\tfrac{E_t}{M} + \tfrac{G^2}{M}\Big). \tag{21}$$

Finally take total expectation and use Lemma 1 to control E_t.

B.5. Proof of rates

Convex case (diminishing steps). Assume f is convex and L-smooth, and $\{\rho_t\}$ satisfies $\sum_t \rho_t = \infty, \sum_t \rho_t^2 < \infty$. Using the standard smooth convex inequality $\|\nabla f(x)\|_2^2 \ge 2L\,(f(x) - f^\dagger)$ is *not* valid; instead we use the classical descent analysis via potential $\|\bar{\mu}_t - \mu^\dagger\|_2^2$: for any $\mu^\dagger \in \arg\min f$,

$$\|\bar{\mu}_{t+1} - \mu^\dagger\|_2^2 = \|\bar{\mu}_t - \mu^\dagger\|_2^2 - 2\rho_t\langle \bar{g}_t, \bar{\mu}_t - \mu^\dagger\rangle + \rho_t^2\|\bar{g}_t\|_2^2.$$

Taking the conditional expectation, using convexity $\langle \nabla f(\bar{\mu}_t), \bar{\mu}_t - \mu^\dagger\rangle \ge f(\bar{\mu}_t) - f^\dagger$, and bounding the perturbation $\bar{g}_t - \nabla f(\bar{\mu}_t)$ via Lemma 3 (and E_t via Lemma 1), this yields a standard telescoping bound,

$$\sum_{t=0}^{T-1} \rho_t\,\mathbb{E}[f(\bar{\mu}_t) - f^\dagger] \le \mathcal{O}\big(\|\bar{\mu}_0 - \mu^\dagger\|_2^2\big) + \mathcal{O}\left(\frac{1}{1-\Delta}\sum_{t=0}^{T-1}\rho_t^2\right) + \mathcal{O}\left(\sum_{t=0}^{T-1}\rho_t^2\right).$$

Dividing by $\sum_{t<T}\rho_t$ gives

$$\mathbb{E}[f(\bar{\mu}_T) - f^\dagger] = \mathcal{O}\left(\frac{\|\bar{\mu}_0 - \mu^\dagger\|_2^2}{\sum_{t<T}\rho_t}\right) + \mathcal{O}\left(\frac{1}{1-\Delta}\cdot\frac{\sum_{t<T}\rho_t^2}{\sum_{t<T}\rho_t}\right). \tag{22}$$

For $\rho_t = \rho_0/(t+\tau)$, we have $\sum_{t<T}\rho_t = \Theta(\log T)$ and $\sum_{t<T}\rho_t^2 = \Theta(1)$, yielding the customary $\tilde{\mathcal{O}}((1-\Delta)^{-1}T^{-1})$-type behavior up to log factors; for piecewise-constant schedules, Eq. (22) recovers the stated $\mathcal{O}((1-\Delta)^{-1}T^{-1})$ scaling over each stage.

Strongly convex case (constant step). Assume f is μ_c-strongly convex and L-smooth. Then

$$\|\nabla f(x)\|_2^2 \ge 2\mu_c\,(f(x) - f^\dagger) \quad \text{for all } x. \tag{23}$$

Choose a constant step $\rho_t \equiv \rho$ small enough so that $1 - \frac{L\rho}{2} \ge \frac{1}{2}$. Taking the expectation in Eq. (21), using Eq. (23), and bounding $\mathbb{E}[E_t]$ by Lemma 1 gives

$$\mathbb{E}[f(\bar{\mu}_{t+1}) - f^\dagger] \le (1 - \rho\mu_c)\,\mathbb{E}[f(\bar{\mu}_t) - f^\dagger] + \mathcal{O}\left(\frac{L\rho^2}{1-\Delta}\right) + \mathcal{O}\left(\frac{L\rho^2 G^2}{M}\right).$$

Unrolling the recursion yields linear convergence to a neighborhood of radius $\mathcal{O}(\rho/(1-\Delta))$ (and proportional to G^2 when making the constants explicit), matching the main-text statement.

Geometry constant κ_m. When relating contraction/descent constants measured in the Euclidean μ-norm to those induced by the EF metric, Eq. (14) implies condition-number-type factors. One convenient choice is $\kappa_m = u_F/\ell_F$, which captures the relative scaling between the Fisher geometry and the Euclidean norm used in defining L and μ_c. This is the quantity appearing implicitly in the step-size restriction and the linear rate constant. □

C. Experiment details

Experiments on CIFAR-100 We use ResNet-50 with BatchNorm replaced by GroupNorm. Centralized training uses SGD (initial learning rate of 0.1 and cosine annealing), batch size 128, weight decay 5×10^{-4} for 250 epochs. Gossip-SGD, SGP, GT-SGD, and QGM use the same optimizer and schedule as the centralized run. KL-consensus Adam uses Adam (initial learning rate of 0.01) with the same cosine schedule.

We visualize per-client label distributions for 8 clients (Fig. 6) and 16 clients (Fig. 7) under varying Dirichlet concentration α and imbalance δ. Lower α increases non-i.i.d. severity. With $\alpha = 0.01$, many classes appear on only one or two clients, and other clients see no examples of those classes. Lower δ increases size imbalance. With $\delta = 0.02$, client 0 holds only about 2% as many samples as each other client.

Experiments on Kvasir-SEG We use a U-Net with a ResNet-34 encoder, again replacing BatchNorm with GroupNorm. Centralized training uses SGD (an initial learning rate of 0.01 and cosine annealing), batch size 16, weight decay 1×10^{-4}, for 100 epochs. Gossip-SGD, SGP, GT-SGD, and QGM match the centralized settings. KL-consensus Adam uses Adam (initial learning rate of 0.005) with the same cosine schedule.

Computation and communication cost We compare the computation and communication cost of Euclidean baselines with our KL-consensus Adam in Table 6. We note that our method keeps the same communication budget of SGP/QGM and uses half the communication of GT-SGD.

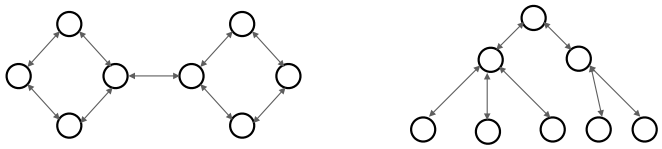

Figure 5. Communication graphs with 8 clients other than rings. Left: two rings with a gateway. Right: hierarchical tree.

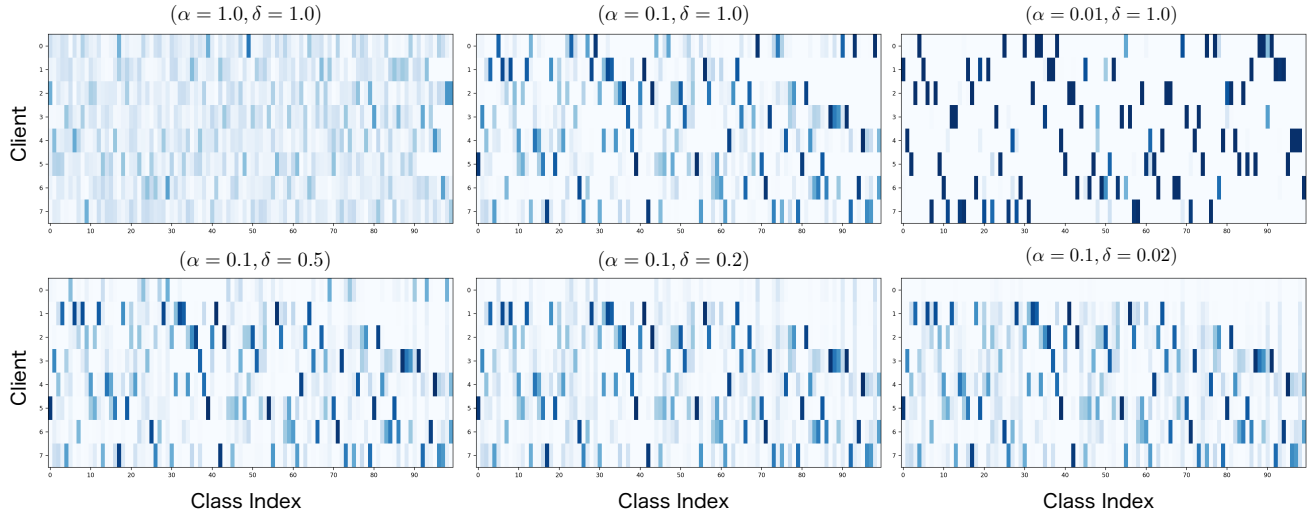

Figure 6. Per-client label distribution of CIFAR-100 for 8 clients.

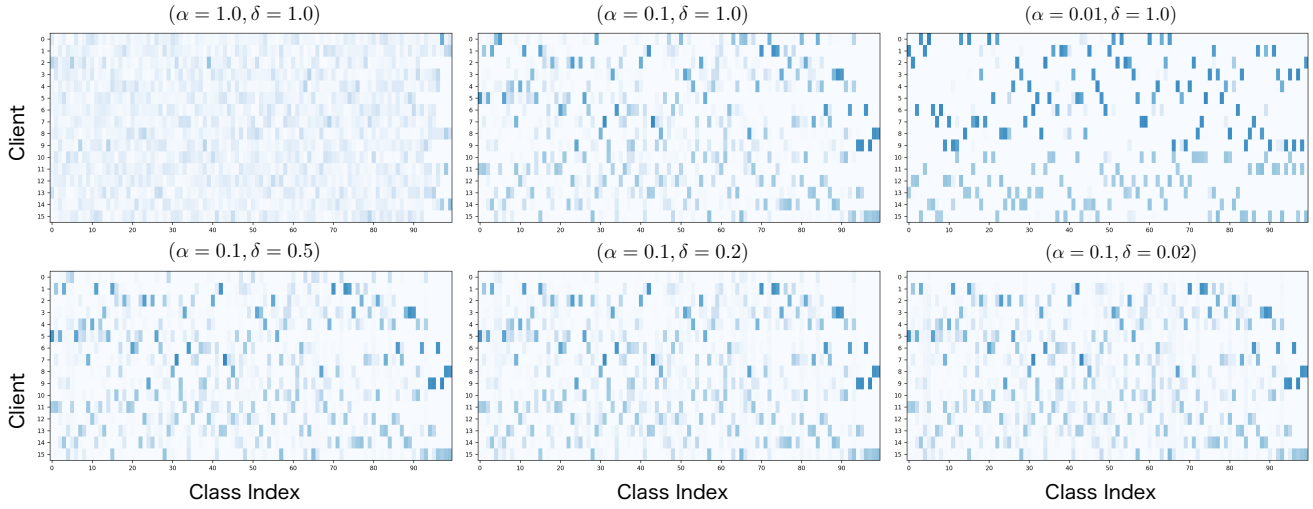

Figure 7. Per-client label distribution of CIFAR-100 for 16 clients.

Table 6. Per-round cost profile (per client), where d is the number of parameters. The communication is counted as d-dimensional vectors sent to each neighbor.

Method	Local compute	Extra memory	Communication round
SGP	$O(d)$	none	$1 \times d$
GT-SGD	$O(d)$	$1 \times d$ (tracker)	$2 \times d$
QGM	$O(d)$	$1 \times d$ (momentum)	$1 \times d$
KL-consensus Adam (ours)	$O(d)$	$2 \times d$ (m, v)	$1 \times d$